Simeoni, I., Stephens, J. C., Hu, F., Deevi, S. V. V., Megy, K., Bariana, T. K., Lentaigne, C., Schulman, S., Sivapalaratnam, S., Vries, M. J. A., Westbury, S. K., Greene, D., Papadia, S., Alessi, M-C., Attwood, A. P., Ballmaier, M., Baynam, G., Bermejo, E., Bertoli, M., ... Turro, E. (2016). A high-throughput sequencing test for diagnosing inherited bleeding, thrombotic, and platelet disorderdisorders. *Blood*, *127*(23), 2791-2803. https://doi.org/10.1182/blood-2015-12-688267

Title: A comprehensive high-throughput sequencing test for the diagnosis of inherited bleeding, thrombotic and platelet disorders

Running title: The ThromboGenomics platform

Author List

Ilenia Simeoni[1,2,3], Jonathan C Stephens[1,2,3], Fengyuan Hu[1-4], Sri Deevi[1,2,3], Karyn Megy[1,2,3], Tadbir K Bariana[5,6], Claire Lentaigne[7,8], Sol Schulman[9], Suthesh Sivapalaratnam[1,2,10], Minka JA Vries[11], Sarah K Westbury[12], Daniel Greene[1,2,3,13], Sofia Papadia[1,2,3], Marie-Christine Alessi[14], Antony P Attwood[1,2,3], Matthias Ballmaier[15], Gareth Baynam[16,17,18,19,20], Emilse Bermejo[21], Marta Bertoli[22], Paul F Bray[23], Loredana Bury[24], Marco Cattaneo[25], Peter Collins[26], Louise C Daugherty[1,2,3], Rémi Favier[27], Deborah L French[28], Bruce Furie[9], Michael Gattens[29], Manuela Germeshausen[15], Cedric Ghevaert[1,2], Anne Goodeve[30], Jose Guerrero[1,2], Daniel J Hampshire[31], Daniel Hart[10], Johan WM Heemskerk[32], Yvonne Henskens[11], Marian Hill[33], Nancy Hogg[34], Jennifer D Jolley[35], Walter H Kahr[36], Anne M Kelly[37], Ron Kerr[38], Myrto Kostadima[1,2,3], Shinji Kunishima[39], Michele P Lambert[28,40], Ri Liesner[37], Jose Lopez[41], Rutendo Mapeta[1,3], Mary Mathias[37], Carolyn M Millar[7,8], Amit Nathwani[6], Marguerite Neerman-Arbez[42], Alan T Nurden[43], Paquita Nurden[43], Maha Othman[44], Kathelijne Peerlinck[45], David J Perry[29], Pawan Poudel[46], Pieter Reitsma[47], Matthew Rondina[48], Peter A Smethurst[1,35], William Stevenson[49], Artur Szkotak[50], Salih Tuna[1,2,3], Christel van Geet[45] Deborah Whitehorn[1,3], David A Wilcox[51], Bin Zhang[52], Shoshana Revel-Vilk[53], Paolo Gresele[24], Daniel Bellissimo[54], Christopher J Penkett[1,2,3], Michael A Laffan[7,8], Andrew D Mumford[12,55], Augusto Rendon[1,56], Keith Gomez[5,6,*], Kathleen Freson[45,*], Willem H Ouwehand[1,2,3,57,*], Ernest Turro[1,2,3,13,*]

Affiliate Institutions

[1] Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, United Kingdom. [2] NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, United Kingdom. [3] NIHR BioResource - Rare Diseases, Cambridge University Hospitals, Cambridge Biomedical Campus, Cambridge, United Kingdom. [4] Babraham Institute, Cambridge, United Kingdom. [5] The Katharine Dormandy Haemophilia Centre and Thrombosis Unit, Royal Free London NHS Foundation Trust, London, United Kingdom. [6] Department of Haematology, University College London Cancer Institute, London, United Kingdom. [7] Centre for Haematology, Hammersmith Campus, Imperial College Academic Health Sciences Centre, Imperial College London, London, United Kingdom. [8] Imperial College Healthcare NHS Trust, London, United Kingdom. [9] Beth Israel Deaconess Medical Centre, Harvard Medical School, Boston, MA. [10] Barts Health NHS Trust, London, United Kingdom. [11] Department of

Biochemistry, Maastricht University, The Netherlands. [12] School of Cellular and Molecular Medicine, University of Bristol, Bristol, United Kingdom. [13] Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, Cambridge Biomedical Campus, Cambridge, United Kingdom. [14] Institut National de la Santé et de la Recherche Médicale U626, Faculté de Médecine, Marseille, France. [15] Pediatric Hematology and Oncology, Hannover Medical School, Hannover, Germany. [16] School of Paediatrics and Child Health, The University of Western Australia, Crawley, Australia. [17] Western Australia Register of Developmental Anomalies, King Edward Memorial Hospital, Western Australia, Subiaco, Australia. [18] Office of Population Health Genomics, WA Department of Health, Western Australia, East Perth, Australia. [19] Institute of Immunology and Infectious Diseases, Murdoch University, Western Australia, Murdoch, Australia. [20] Telethon Kids Institute, Western Australia, Subiaco, Australia. [21] Hematological Research Institute, National Academy of Medicine, Buenos Aires, Argentina. [22] San Pietro Hospital, Rome, Italy. [23] Thomas Jefferson University, Jefferson Medical College, Philadelphia, Pennsylvania, PA. [24] Department of Internal Medicine, Section of Internal and Cardiovascular Medicine, University of Perugia, Perugia, Italy. [25] Dipartimento di Scienze Della Salute, Università Degli Studi di Milano, Unità di Medicina 3, Azienda Ospedaliera San Paolo, Milano, Italy. [26] Arthur Bloom Haemophilia Centre, Institute of Infection and Immunity, School of Medicine, Cardiff University, United Kingdom. [27] Assistance Publique - Hôpitaux de Paris, Armand Trousseau Children Hospital, Paris, Inserm U1170, Villejuif, France. [28] Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA. [29] Department of Haematology, Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, United Kingdom. [30] University of Sheffield, Medical School, Beech Hill Road, Sheffield, United Kingdom. [31] Department of Infection, Immunity & Cardiovascular Disease, Haemostasis Research Group, Department of Cardiovascular Science, Faculty of Medicine, Dentistry and Health, University of Sheffield, Sheffield, United Kingdom. [32] Department of Biochemistry, Cardiovascular Research Institute Maastricht, University of Maastricht, Maastricht, The Netherlands. [33] Department of Clinical Pathology, Nottingham University Hospitals NHS Trust, Nottingham City Hospital, Nottingham, United Kingdom. [34] Cancer Research UK London Research Institute: Lincoln's Inn Fields, London, United Kingdom. [35] Components Development Laboratory, NHS Blood and Transplant, Cambridge, United Kingdom. [36] The Hospital for Sick Children, Toronto, Ontario, Canada. [37] Department of Haematology, Great Ormond Street Hospital for Children NHS Trust, London, United Kingdom. [38] Department of Haematology, Ninewells Hospital, Dundee, United Kingdom. [39] Clinical Research Center, National Hospital Organization Nagoya Medical Center, Nagoya, Japan. [40] Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA. [41] Puget Sound Blood Center, Research Institute, Seattle, WA. [42] Department of Genetic Medicine and Development, University Medical Centre, Geneva, Switzerland. [43] Institut Hospitalo-Universitaire LIRYC, PTIB, Hôpital Xavier

Arnozan, Pessac, France. [44] Department of Biomedical and Molecular Sciences, School of Medicine, Queens University, Kingston Ontario, Canada. [45] Department of Cardiovascular Sciences, Center for Molecular and Vascular Biology, University of Leuven, Belgium. [46] Institute of Cancer Research, London, United Kingdom. [47] Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, The Netherlands. [48] University of Utah School of Medicine, Eccles Institute of Human Genetics, Utah, UT. [49] Sydney Medical School, University of Sydney, Sydney, Australia. [50] Laboratory Medicine and Pathology, University of Alberta, Edmonton, Canada. [51] Children's Hospital of Wisconsin Research Institute, Medical College of Wisconsin, Milwaukee, WI. [52] Genomic Medicine Institute, Cleveland Clinic Lerner Research Institute, Cleveland, OH. [53] Pediatric Hematology/ Oncology Department, Hadassah- Hebrew University Medical Center, Jerusalem, Israel. [54] Clinical Genomics Laboratory, University of Pittsburgh, PA. [55] School of Clinical Sciences, University of Bristol, United Kingdom. [56] Genomics England Ltd, London, United Kingdom. [57] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom.

* Authors contributed equally to the study

Corresponding Authors:

Dr Keith Gomez, The Katharine Dormandy Haemophilia Centre and Thrombosis Unit, Royal Free London NHS Foundation Trust, Royal Free Hospital, Pond Street, London, NW3 2QG, UK;
Tel +44 (0)20 7830 2068; email k.gomez@ucl.ac.uk

Professor Willem H Ouwehand, University of Cambridge, Department of Haematology, Long Road, Cambridge, CB2 0PT, UK;
Tel: +44 (0)1223 58 8037/8018, Fax: +44 (0) 1223 588155
email: who1000@cam.ac.uk

Key points

- We have developed a targeted sequencing platform covering 63 genes linked to heritable bleeding, thrombotic and platelet disorders.
- The ThromboGenomics platform provides a sensitive genetic test to obtain molecular diagnoses in patients with a suspected etiology.

ABSTRACT

Inherited bleeding, thrombotic and platelet disorders (BPDs) are diseases affecting approximately 300 individuals per million births. With the exception of haemophilia and von Willebrand disease patients, a molecular analysis for patients with a BPD is often unavailable. Many specialised tests are usually required to reach a putative diagnosis and they are typically performed in a step-wise manner to control costs. This approach causes delays and a conclusive molecular diagnosis is often never reached which can compromise treatment and impede rapid identification of affected relatives. To address this unmet diagnostic need, we designed a high-throughput sequencing (HTS) platform targeting 63 genes relevant for BPDs. The platform can call single nucleotide variants (SNVs), short insertions/deletions (indels) and large copy number variants (CNVs), though not inversions, which are subjected to automated filtering for diagnostic prioritization, resulting in an average of 5.34 candidate variants per individual. We sequenced 159 and 137 samples respectively from cases with and without previously known causal variants. Among the latter group, 61 cases had clinical and laboratory phenotypes indicative of a particular molecular etiology while the remainder had an *a priori* highly uncertain etiology. All previously detected variants were recapitulated and, when the etiology was suspected but unknown or uncertain, a molecular diagnosis was reached in 56 of 61 and only eight of 76 cases, respectively. The latter category highlights the need for further research into novel causes of BPDs. The ThromboGenomics platform thus provides an affordable DNA-based test to diagnose patients suspected of having a known inherited BPD.

INTRODUCTION

HTS of genomic DNA is being introduced into clinical practice[1] and broadly falls into two categories: whole genome sequencing (WGS)[2] and targeted sequencing of pre-specified regions of the genome by means of probe-based capture. These regions may include all exons (whole exome sequencing (WES)) and be sequenced to moderate depth or they may comprise a much smaller fraction of the genome and be sequenced to high depth[3]. Capture probes targeting regions that have been widely studied and implicated in a group of rare heritable disorders can turn HTS into a valuable tool for their affordable diagnosis.

In this work, we focus on the diagnosis of rare heritable BPDs. Previously, we have defined a BPD case as a patient having abnormal platelet count, volume, morphology or function, or with a tendency to bleed abnormally[4]. The abnormal phenotypes must furthermore be judged to have a genetic basis, thereby ruling out diseases that may have been acquired or thought to be caused by exposure to known environmental risk factors. For the purposes of this paper, we also include patients with an abnormal tendency to thrombus formation in our definition. Currently 90% of BPD cases who do not have hemophilia[5] or von Willebrand disease[6,7] (VWD) never receive a conclusive molecular diagnosis due to the unavailability of affordable genetic tests[8]. Hence treatment is compromised in some cases and the rapid identification of affected relatives may be impeded.

The aims of the ThromboGenomics project are to develop a multi-gene HTS platform for the diagnosis of BPDs, to deposit knowledge about novel pathogenic variants in a sustainable and freely available database and to leverage systematic Human Phenotype Ontology (HPO)-term based coding of patient phenotypes[4] to improve our understanding of genotype-phenotype correlations in BPDs. To deliver the project to high scientific and ethical standards a global ThromboGenomics network of clinicians and researchers

with expertise in BPDs (supplemental Figure 1) was formed. Currently, the ThromboGenomics HTS platform can detect variants in the exonic fraction of 63 BPD genes and many of their introns and untranslated regions (UTRs). Multiplexing allows the sequencing of DNA samples from 24 cases simultaneously. A custom capture is used instead of WES because it provides deeper coverage of the regions of interest for a given number of sequencing reads and allows a higher grade of multiplexing, thereby reducing the cost per patient diagnosed.

Here we describe the technical performance of the ThromboGenomics platform and its accuracy for detecting causal variants in a curated set of transcripts in 63 BPD genes. We sequenced 300 samples (Figure 1), of which 260 are unrelated, drawn from four groups of individuals having: diagnostic abnormalities on laboratory assays with previously ascertained pathogenic variants (*known* group; n=159), phenotypes strongly indicative of a particular disorder on the basis of laboratory abnormalities, but without knowledge of causal variants (*suspected* group; n=61), phenotypes that could not be matched to any known BPD because the laboratory assays were either normal or not diagnostic of an established disorder (*uncertain* group; n=76) and four samples from *unaffected* relatives.

We developed a variant filtering procedure, assessed the platform's reproducibility, used HPO-based prioritisation of candidate variants and reviewed the quality of the variant pathogenicity literature. Finally, we discuss the rules used by the multidisciplinary team (MDT) to review the sequencing results and generate reports.

METHODS

Enrollment, gene and transcript selection, platform design and sequencing

Individuals were enrolled in 13 different countries. The gene list was established through discussion with experts in BPDs and taking into account the quality of the evidence in the peer-reviewed literature supporting the claim to pathogenicity of variants in each gene, such as the number of unrelated individuals carrying the variants and functional validation of their effects in vitro. Transcripts were selected by experts for each gene, where available. The platform was designed to target all exonic and many intronic and UTRs of BPD genes. For further details on gene and transcript selection, platform design, sample preparation and sequencing, see supplemental Materials.

Clinical bioinformatics

The reads in the de-multiplexed paired-end FASTQ files are processed as described in the supplemental Materials. Briefly, reads are aligned using BWA[9] 0.7.10. Then, SNVs and indels are called using GATK[10] 3.3 HaplotypeCaller and CNVs are called using ExomeDepth[11] 1.1.5. As it is not possible to call inversions and complex structural rearrangements accurately with capture technology, they are not called.

We infer gender using two statistics based on sequence reads aligned to well-covered target regions (>95% of samples covered at 20X): the ratio between heterozygote and non-reference homozygote genotypes (het/hom) on the X chromosome sites and the ratio between the median coverage on X and the median coverage on the autosomes (aut/X). The het/hom ratio is computed using heterozygote SNVs with an allele depth between 0.3 and 0.7 to guard against errors.

In order to estimate ethnic background, we project standardised genotypes onto the first two principal components obtained from the standardised

genotypes of 2,504 HapMap[12] individuals sequenced by the 1000 Genomes project[13]. We use SNVs falling within well-covered target regions (>95% of samples covered at 20X), having a minor allele frequency (MAF) >0.02 in 1000 Genomes and pruned with PLINK[14] to ensure that r^2<0.2 between pairs of SNVs.

We annotate SNVs and indels with their predicted impact against Ensembl 75, presence in HGMD 2015.2 and their MAFs in the ExAC (http://biorxiv.org/content/early/2015/10/30/030338) release 0.3 and 1000 Genomes[13] databases using SnpEff[15] 4.0. If variants are in HGMD[16], then they are retained as long as their MAF in ExAC and in 1000 Genomes is <2.5%. Otherwise, variants must have a MAF <0.1% in ExAC and 1000 Genomes, have <4 alternate alleles (as typically found in repetitive regions) and have a predicted moderate or high impact on translation of one of the ThromboGenomics transcripts according to SnpEff. In this latter group, we also allow through variants predicted to affect splice regions if they do not fail quality control in ExAC. Finally, any variants with a MAF >10% within the entirety of the ThromboGenomics data set are filtered out to remove potential systematic artefacts. The filtering criteria were informed by the variants in the known category of samples but applied universally.

Our approach allows us to retain confidently called pathogenic variants which are regulatory or moderately common if they are already known to be pathogenic. Such variants include, for instance, the regulatory non-coding SNVs in *RBM8A*[17,18] responsible for Thrombocytopenia Absent Radius (TAR) syndrome in the presence of a loss-of-function variant on the alternate haplotype, moderately common variants in *VWF* linked to reduced levels of the VWF protein[19-21] and the *F5* Leiden[22] variant.

HPO-based prioritisation of variants

We compute the semantic similarity *S* of a case's HPO-coded phenotype to that of the HPO profiles of a BPD gene using the "best match average" metric[23] and Lin's[24] measure of similarity between terms:

$$\begin{aligned}
\mathrm{IC(t)} &= -\log\left(\mathrm{freq}(t)\right), \\
\mathrm{MICA}(t_1, t_2) &= \max_{t \in \mathrm{anc}(t_1) \cap \mathrm{anc}(t_2)} \mathrm{IC}(t), \\
s(t_1, t_2) &= \frac{2 * \mathrm{MICA}(t_1, t_2)}{\mathrm{IC}(t_1) + \mathrm{IC}(t_2)}, \\
S'(x_1, x_2) &= \frac{1}{|x_1|} \sum_{t_1 \in x_1} \max_{t_2 \in x_2} s(t_1, t_2) \\
S(x_1, x_2) &= \frac{S'(x_1, x_2) + S'(x_2, x_1)}{2}
\end{aligned}$$

where freq(*t*) is the frequency of term *t* in OMIM, anc(*t*) refers to union of term *t* and its ancestors in the HPO and (x_1, x_2) refers to the pair of sets of HPO terms being compared.

Web application for variant assessment

Variant calls and phenotypes are visualised in the Sapientia web application (Congenica Inc., Cambridge, UK) during MDT meetings. Sapientia displays variant information such as predicted effect, MAFs in reference cohorts (e.g. ExAC, UK10K[25]) and links to external resources (e.g. HGMD, ClinVar, PubMed), as well as showing case data such as phenotype information in the form of HPO terms. The web application allows the MDT to annotate each variant with respect to its predicted contribution to the disease phenotype and its likely pathogenicity. Following international guidelines[26], variants are marked as pathogenic (high impact or previously observed in at least four unrelated cases with a similar pathology), likely pathogenic (previously observed in <4 unrelated cases with a similar pathology) or as having unknown clinical significance. Considerations such as predicted impact (e.g. missense or loss of function) and conservation/pathogenicity scores are used

to inform how variants are categorised in the context of the observed phenotype. After deliberation the MDT produces a research report for the referring clinician including information about variants declared pathogenic or likely pathogenic but not usually about variants of unknown significance (VUS) (supplemental Figure 2). The typical overall turnaround time from sample submission to the production of a research report is less than 16 weeks.

RESULTS

Coverage profile of the ThromboGenomics platform

We have assessed the sequencing coverage profile of the ThromboGenomics platform using data from 300 samples. The mean exonic coverage along the 58 autosomal genes (comprising 228,863 bp; supplemental Table 1) across samples from individuals of both sexes was on average 1,178 (range 123 to 2,356) (Figure 2A). The mean fraction of exonic bases covered at 20X and 50X was 0.993 and 0.989 respectively (Figure 2B). We have produced individualised coverage profiles for each gene on the platform showing that virtually all exonic regions of the ThromboGenomics transcripts are covered sufficiently for sensitive variant calling (supplemental File 1). The profile of *ITGA2B* encoding αIIb of the major platelet integrin is shown in Figure 2C as an example. A small number of short regions that may potentially suffer from low coverage are also highlighted in 19 genes (supplemental Table 2). Overall, 613 bp overlapping coding regions and only 44 out of the 8,294 HGMD variants (0.53%) for the 63 genes were covered <20X in 95% of samples. Of these, 22 variants lie in exon 26 of *VWF*, which is perfectly homologous with part of the von Willebrand pseudogene 1 and 20 HGMD variants are in the highly GC-rich exon 2 of *GP1BB*. However, a change in the polymerase enzyme used in the library preparation improved the coverage on this (supplemental Figure 3) and other GC-rich regions such that all HGMD variants except for those on exon 26 of *VWF* could subsequently be called with confidence.

Sample identity assurance

The MDT ensures that the gender (Figure 3A) and ethnic background (Figure 3B), including admixture, inferred from the genotype data match the information provided by the clinical care team. Caucasoid individuals are over-represented in the large collections of samples used for allele frequency based filtering. Consistent with this notion, individuals with non-European ancestry (particularly East Asian or African) tend to have more candidate variants (after filtering) than European or South Asian individuals (Figure 3B).

Platform reproducibility

The ThromboGenomics platform increases the total length of genome sequenced per case over 80-fold compared to a typical Sanger sequencing based test[27], which raises the concern that a false diagnosis may arise from a spurious match between phenotype and a falsely called genotype. It is not feasible to verify all variants (including non-pathogenic variants) systematically using an alternative gold standard genotyping method and as such we cannot obtain a direct estimate of the false discovery rate (FDR) of the ThromboGenomics platform. However, any susceptibility towards spurious genotyping calls would likely manifest in low concordance rates between the variants obtained from different sequencing runs of the same sample. We thus assessed whether FDR is reasonably controlled by sequencing six DNA samples in two separate runs and comparing the candidate variants obtained between replicates. We found that every one of the 22 candidates called in either replicate was found in both replicates for all six pairs of samples. These results indicate that the library preparation, sequencing, variant calling and variant filtering altogether produce reproducible results and the risk of erroneous diagnoses due to falsely called variants is negligible.

Overall, 75 samples had at least one of the CNV calls produced by the ExomeDepth algorithm comprising 47 deletions and 28 amplifications. To assess the dependability of these calls, we focused on the 29 heterozygous

autosomal deletions, found in 44 different samples, expected to be significantly depleted of heterozygous SNV calls under the hypothesis that the samples are truly haploid. In 82% of cases, no heterozygous SNVs spanned the putative deleted regions at all and the remainder contained between one and three heterozygous SNVs (supplemental Figure 4). Although some of these heterozygote calls may be due to read misalignment, a small proportion of CNV calls are likely the result of artifactual changes in coverage. Nevertheless, the highly significant overall depletion ($p < 10^{-6}$), coupled with a proven sensitivity to identify pathogenic CNVs (see below), indicates that a reasonable balance between sensitivity and specificity is achieved by ExomeDepth in the difficult task of identifying CNVs. Given the symmetric modelling of deletions and amplifications, we expect similar performance in calling duplications.

Performance of variant calling and filtering

The DNA from 300 individuals was sequenced and, across the entire dataset, 20,039 variants were called, of which 520 SNVs, 47 indels and 75 CNVs remained after automated filtering. The mean number of variants per individual before filtering is 2,014.11 and this is reduced to 5.34 candidate causal variants (range 2 to 12) by the filtering procedure. Assuming there are two causal variants per individual and the filtering method classifies variants into candidates and non-candidates for pathogenicity in the worst possible way gives a lower bound for the specificity of variant filtration of 99.73%. On average, the 5.34 candidate variants consisted of 4.74 SNVs, 0.35 indels and 0.25 CNVs (Figure 4). The CNV reads ratios clustered into groups corresponding to different zygosity (Figure 4D).

The candidate variants obtained from 159 individuals in the *known* group were used to assess the sensitivity of the platform. Members of this group had, or shared carrier status with, relatives who had one of 30 different BPDs. They included 19 with *MYH9*-related disorder, 11 with Glanzmann thrombasthenia

and 10 each with TAR syndrome, Wiskott–Aldrich syndrome (WAS), Platelet-type VWD, Fibrinogen deficiency and Autosomal dominant thrombocytopenia. A further eight and six individuals had Gray platelet and Hermansky-Pudlak syndromes, respectively. The remaining 65 individuals in this group carried variants underlying other BPDs (Table 1). The previously known variants included 119 SNVs, 19 indels and seven large deletions in 37 different BPD genes. Of the 138 SNVs and indels, 102 (73.9%) were HGMD pathogenic variants and the remaining 36 variants were deemed to be causal by the clinician who submitted the sample and confirmed as likely pathogenic by the MDT. The longest indel detected in this group was called in DNA samples from two related patients diagnosed with platelet-type von Willebrand disease due to a heterozygous 27 bp in-frame deletion in GP1BA that removes amino acids 459–467. The large deletions varied in size, ranging from at least one exon to entire genes. After initial tuning of the bioinformatics pipeline, all causal variants across known BPD cases were identified after filtering and the genotypes matched the previously determined zygosity of the corresponding samples. Thus, the ThromboGenomics platform has an empirical sensitivity based on these 159 samples harbouring 145 causal variants (which do not include inversions) of 100% to detect known causal variants in BPD genes.

Yield of the ThromboGenomics platform for cases with a *suspected* etiology

We evaluated the utility of the platform in a clinical diagnostic setting by processing 61 samples from the *suspected* group, of which 52 are unrelated, using the same bioinformatics parameter settings as above. This group includes 28, 24 and nine cases with a coagulation factor, platelet or thrombotic disorder, respectively. The called variants were reviewed by the MDT in the context of the HPO terms annotated for each case[29]. In all but five of the 61 cases (91.8%; 90.4% for probands only), the MDT identified pathogenic or likely pathogenic variants that fully or partially explained the disease phenotype (Table 1). Overall, we identified 29 pathogenic and 28

likely pathogenic variants, comprising 44 SNVs, 13 indels and one duplication, of which 28 variants were novel and the remaining ones had previously been deposited in HGMD as BPD-causing variants.

A noteworthy example concerns two cases from the same pedigree who were coded with the HPO terms “Impaired platelet aggregation”, “Spontaneous, recurrent epistaxis” and “Intramuscular hematoma” (Figure 5A). During analysis, all SNVs and indels that passed filtering were determined to be VUS. However, a duplication spanning the entirety of the *PLAU* gene was uncovered in both pedigree members (Figure 5B), hence a positive diagnosis was reached for Québec platelet disorder[28]. This example highlights the value of the ThromboGenomics platform because the standard laboratory marker of this condition, platelet aggregometry, is by no means conclusive.

The unexplained cases included two with suspected Protein S deficiency, one with suspected Bernard-Soulier syndrome (BSS) and one with suspected type 1 VWD. We have not yet been able to determine whether this was due to a lack of sensitivity (e.g. because of a complex rearrangement that cannot be detected by targeted sequencing) or because the cases had causative variants in relevant regulatory elements or in other trans-acting genes. In type 1 VWD pathogenic variants are found in *VWF* in only ~50% of cases[29] using Sanger sequencing. Alternatively the three cases may have an acquired BPD.

We note that several samples in the *suspected* group were submitted after a negative result had been returned by Sanger sequencing in genes thought to harbour causal variants. In five cases, this result was overturned by the detection of causal variants by ThromboGenomics sequencing, which was subsequently confirmed in a second round of Sanger sequencing. A noteworthy example concerns two members in the same pedigree, initially diagnosed with alpha/delta-storage pool disease. The ThromboGenomics test results revealed a mutation in the *RUNX1* gene changing the diagnosis to a

RUNX-1 associated thrombocytopenia with increased risk of acute myeloid leukemia. A further example illustrating how the ThromboGenomics platform can deliver more pertinent information than a standard single gene screening approach relates to a one-year old boy with thrombocytopenia, normal mean platelet volume and multiple hematomas after preterm birth without any other symptoms and having parents with normal platelet counts. Variants were found in both *MYH9* and *WAS* that were coded by the MDT as 'likely pathogenic' and 'VUS', respectively. His mother carries a variant in *WAS* (located on the X chromosome) that was previously described for another male *WAS* patient[30] while his father has a variant in *MYH9* variant that has not been described previously. Further studies are required to determine whether the co-inheritance of the variants in *MYH9* and *WAS* are causal of the lack of effective hemostasis. This example highlights the advantage of an HTS strategy that can identify potential disease-modifying factors, acting in trans, over a single gene analysis strategy.

Yield of the ThromboGenomics platform for cases with a highly *uncertain* etiology

The third group of 76 *uncertain* cases, of which 62 are unrelated, is made up of a mixture of cases with unexplained BPDs that are not suggestive of a particular known pathology. This group mainly comprises patients with clinical bleeding problems, but having normal laboratory coagulation and platelet function tests, platelet storage pool disorder or patients who have had thrombotic events and low protein S levels though with a normal *PROS1* gene. We detected pathogenic or likely pathogenic variants in eight cases, corresponding to a sensitivity of 10.5% (9.68% for probands only). In two cases, the variants uncovered a contributory defect in a coagulation factor that explained the phenotype only partially (e.g. due to reduced levels that did not explain the bleeding). In the remaining six cases, defects explaining the phenotype in full were found in *MYH9*, *PROC, PROS1*, *RUNX1, SERPINC1* and *TUBB1*, including a digenic molecular diagnosis involving a 'likely

pathogenic' variant in *SERPINC1* and another in *PROC*, possibly explaining the thrombotic phenotype observed in this patient.

Negative Sanger sequencing results were overturned in three cases within this group also, demonstrating that the ThromboGenomics platform can outperform Sanger sequencing in terms of sensitivity. However, the vast majority of cases in the *uncertain* group were given a negative result by the MDT, which underscores the need for further research into the molecular etiology of uncharacterised BPDs.

HPO-based prioritisation of candidate variants

We transcribed phenotypes linked to the diseases associated with the 63 BPD genes into HPO terms (supplemental Table 3) and obtained HPO terms describing the phenotypes of a subset of the cases in our collection. To assess the potential utility of HPO methods for prioritising candidate variants, we compared the HPO terms for 109 cases who were previously determined to carry a pathogenic or likely pathogenic variant by the MDT to the HPO profiles linked to the genes in which they carried candidate variants. For example, a case with BSS from the *suspected* group was coded with six HPO terms and subsequently found to carry candidate variants in four genes. The homozygous variant in *GP1BB*, identified independently by the MDT as likely pathogenic, was chosen by the prioritisation algorithm as the top candidate because the HPO profile linked to *GP1BB* was more similar to the HPO profile of the case than the profiles of any of the other three genes in which the case had a candidate variant (Figure 6A). The overall results, shown in Figure 6B, indicate that in 85% (93/109) of cases, the correct gene, as identified by the MDT, scored the highest similarity to the case phenotype out of all the candidate variants ($p < 10^{-6}$). Whenever the top-ranked gene did not correspond to the MDT-designated gene, the difference tended to be smaller than when there was concordance (supplemental Figure 5). Thus, HPO-

based prioritisation offers a promising route towards streamlining the review process by MDTs.

On the reliability of the variant pathogenicity literature

Presence of a candidate variant in HGMD is often considered a strong indicator of pathogenicity subject to a phenotype match between the disease linked with the variant, the patient's phenotype and a consistent mode of inheritance (Table 1). The variants in HGMD belong to different classes depending on whether they are considered disease-causing mutations (labelled "DM" if definitive and "DM?" if the curator had reservations), disease-associated polymorphisms or other types of variants. The variant with dbSNP ID rs139428292 in the 5′ UTR of the *RBM8A* gene is causal of TAR if the alternate *RBM8A* allele harbours a loss of function variant. This UTR variant is listed in HGMD as an "*in vitro* or *in vivo* functional polymorphism" instead of as a cause of TAR. The Factor V Leiden variant rs6025 is listed as a "disease-associated polymorphism with additional supporting functional evidence". Consequently, we use HGMD variants in all classes in our MDT analysis.

The vast majority of HGMD variants in the 63 BPD genes (7,320 out of 8,294) have a MAF that is either zero or undetermined in the 60,706 controls from the ExAC database. However, 140 variants have a MAF >1/1,000, of which 69 are listed as disease-causing (Figure 7). We reviewed the literature for these 140 variants and concluded that there is sufficient evidence supporting a claim to penetrant pathogenicity for only seven variants (supplemental Table 4). Historically, assignment of pathogenicity has sometimes been based on publications of variants in small pedigrees without large numbers of control samples or supporting biochemical or cell biology data, such as expression studies. We considered such variants to be of unproven pathogenicity in accordance with current standards. The genes *F5*[31,32], *F8*[33,34], *F11*[35], *PROS1*[35], *FLNA*[36], *THBD*[37,38], *VWF*[39] and *WAS*[40] had the highest rates of

these doubtfully annotated variants, with counts ranging between three (*F5, F11, THBD* and *WAS*) and 17 (*VWF*). By way of example, for *MYH9*, a methionine1651threonine is classified in HGMD as DM. However, the MAF is 1.29 in 1,000 in ExAC and therefore it cannot underlie a high-penetrance autosomal dominant disorder. Indeed, the authors[41] reporting this mutation observed it in a single pedigree in which two cases and no unaffected relatives were screened and decided that its absence in a mere 45 control samples was sufficient to infer causality for the child's Alport syndrome and the mother's hearing loss. Regarding *F8*, the doubtful DM variants are typically in the B-domain and may influence *F8* levels but are unlikely to be causal of haemophilia A. With the above considerations in mind, the ThromboGenomics MDT critically assesses the evidence supporting claims to pathogenicity of each candidate variant in the context of allele frequencies in the major variant databases, even for variants that are present in HGMD.

DISCUSSION

We have described a comprehensive and cost-effective strategy for the diagnosis of BPDs. The HTS platform and accompanying processing and filtering methods have high sensitivity (100% based on 159 samples) to detect and shortlist causal variants (SNVs, indels and CNVs) when the variants are known to be in a BPD gene on the ThromboGenomics platform. When the phenotype is strongly indicative of the presence of a particular disease etiology but the variants are unknown, sensitivity remains high (>90% based on 61 samples). Our variant filtering approach has high specificity (>99.5%) as it greatly reduces the number of candidates that require consideration by the MDT and, as we have shown, HPO-based prioritisation methods may reduce the burden on MDTs even further by highlighting pathogenic or likely pathogenic variants as the top candidate in about 85% of cases. Sanger results have been overturned by results obtained by HTS and the CNV-calling pipeline compares favourably with other assays such as multiplex ligation-dependent probe amplification (MLPA[42]). In order to facilitate interpretation of

the genetic data, variants are annotated against clinically relevant transcripts, which were selected by experts and deposited in the LRG public reference database for use by clinical genetics laboratories.

At MDT meetings, we assume that truly pathogenic variants in BPD genes generally have a MAF <1/1,000 for autosomal recessive disorders and are likely much rarer for X-linked and dominant disorders. Decisions used to reach a diagnosis are guided by MAFs in major reference databases and data extracted from the literature, which has been deposited in the HGMD database. We have shown that variants in all HGMD classes must be considered as potentially pertinent yet 140 variants have a MAF in ExAC >1/1,000, only four of which are established as pathogenic and penetrant, while the others either exert small effects or have uncertain clinical significance. The *VWF* gene, for example, which has an open reading frame length in the top 1% of the genome-wide distribution, has 17 variants labelled as pathogenic or likely pathogenic with a MAF>1/1000 in ExAC. Reasons for this include low control sample numbers and an over-reliance on *in vitro* function tests, pathogenicity prediction algorithms and crystallography data. Given the good performance of the sequencing, variant calling and filtering procedures, the specificity of the ThromboGenomics test as a whole must be determined in large part by the rate at which non-pertinent variants are falsely declared pathogenic. Although we have not been able to measure specificity directly, careful adjudication of the results by the MDT along the lines we have discussed should ensure that false positive reports are rare.

Overall, we have identified 204 distinct pathogenic or likely pathogenic variants, of which 8 are CNVs and 64 are absent from HGMD. The 73 cases for whom no conclusive diagnosis could be reached will be considered for inclusion in the 100,000 Genomes projects to be analysed by WGS. Aggregating these cases with the current set of approximately 1,000 BPD cases already analysed by WES or WGS will improve power to identify novel

causes of BPDs. Meanwhile, the on-going work of the ThromboGenomics project will improve the catalogues of pathogenic variants for known BPD genes to aid future diagnoses. However, as the ThromboGenomics platform cannot identify inversions and ~45% of severe Haemophilia A cases are due to inversions, a simple PCR-based test can be performed to exclude them prior to HTS.

The clinical importance of an affordable HTS test to patient care should not be underestimated. For example, in the UK, sequencing of only *HPS1* and *HPS3* genes for patients with a suspected diagnosis of HPS is reimbursed. However, the precise genetic diagnosis of HPS cases is clinically relevant because those with causal variants in the *HPS*1, 2 and 4 genes may develop lung fibrosis, which requires monitoring, whilst this is not the case for variants in the remaining six HPS genes. Furthermore, the identification of variants in genes like *RUNX1* and *ETV6*, associated with heightened risk of malignancy, would allow patients at risk to benefit from surveillance.

During the validation period of the ThromboGenomics platform, 13 new BPD genes have been reported and these 13 and a further 25 genes (supplemental Table 5) for cerebral small vessel disease, hereditary haemorrhagic telangiectasia, arteriovenous malformations and pulmonary arterial hypertension have been included in the next version of the ThromboGenomics platform. For the version reported in this manuscript we used one capture reaction for every two samples and multiplexed 24 samples on a single HiSeq lane. With improved reagents and protocols for multiplexing and the substantial increase in the number of reads per HiSeq lane the capture of at least 4 samples per reaction will be feasible in the near future, whilst maintaining excellent coverage. As a result of this modification the cost per sample tested can be reduced further.

HTS-based tests are rapidly becoming routine in clinical practice and the ThromboGenomics platform is an example of this transformation. The aim is for the ThromboGenomics test, available through www.thrombogenomics.org.uk, to become the first choice for haemostasis and thrombosis physicians and haematologists requiring a molecular diagnosis for BPD cases. This platform and the underpinning principle of freely accessible expert knowledge about genes, transcripts and causal variants and the approach of using HPO terms for coding phenotype can be used by reference laboratories to reduce the diagnostic delay in reaching a conclusive molecular diagnosis for BPD patients. We believe that, by facilitating provision of a definitive diagnosis, our platform will bring substantial benefits to the estimated 2 million BPD cases worldwide.

Acknowledgments

We thank Roche NimbleGen and Beckman Coulter for their support in the initial stages of this project, Congenica Inc. for adapting the Sapientia software to the needs of the MDT and Jo Westmoreland from the MRC Laboratory for Molecular Biology Visual Aids group for providing the world map picture to represent the ThromboGenomics network. This study, including the enrolment of cases, the sequencing and analysis received support from the NIHR BioResource – Rare Diseases. The NIHR BioResource is funded by the National Institute for Health Research (NIHR; http://www.nihr.ac.uk). Research in the Ouwehand laboratory is also supported by grants from Bristol Myers Squibb, British Heart Foundation, British Society of Haematology, European Commission, Medical Research Council (MRC), NIHR and Wellcome Trust; the laboratory also receives funding from NHS Blood and Transplant (NHSBT). The clinical fellows received funding from the MRC for C.L. and S.K.W., the NIHR – Rare Diseases Translational Research Collaboration for S. Sivapalaratnam and the British Society for Haematology and NHSBT for T.K.B.

Authorship

Contribution: I.S. developed and validated the ThromboGenomics platform, processed samples, sought transcript curators, fostered and coordinated the growth of the ThromboGenomics community, coordinated MDT meetings and wrote the paper; J.C.S. performed experiments; F.H., S.D., and S.T. performed data analysis; K.M. maintained the database of transcripts and contributed to MDT meetings; T.K.B., C.L. and S.K.W. are part of the clinical care team and curated genes; T.K.B., C.L., S.K.W., S.S. and M.J.A.V. provided HPO-coding of BPDs; D.G. performed HPO analysis and S.P. supervised Ethics documentations, T.K.B., C.L., S.K.W., S.S., S. Sivapalaratnam, M.J.A.V., M-C.A., M.B., G.B., E.B., M. Bertoli, L.B., P.C., R.F., B.F., M.G., M. Germeshausen, D.P.H., Y.M.C.H., A.M.K., R.K., M.P.L, R.L., M.M., C.M.M., P.N., M.O., D.J.P., M.R., A.S., VG.C., C.V-G., S.R.V, P.G., M.A.L., A.D.M., K.G., K.F. provided samples; A.P.A. developed database for phenotype collection and HPO coding; P.F.B., D.B., K.F., W.H.O., A.G., M.P.L. and P.R. are Co-chairs of the Genomics in Thrombosis and Hemostasis ISTH-SCC Committee; M.C., D.L.F, C.G., J.G., D.J.H., J.W.M.H, M.H, N.H., W.H.K., S.K., J.L., A.N., M.N-A., A.T.N, M.O., P.R., P.A.S, W.S., D.A.W., B.Z., P.G., D.B., A.D.M., K.F. curated the BPD genes; L.C.D. maintained gene list; J.D.J., R.M. and D.W extracted DNA, K. P. enrolled cases in Belgium; P.P. helped preparing documentation for gene curation, C.J.P managed the bioinformatics pipeline, A.R. developed the initial probe design, K.G. chaired the MDT meetings; K.G., K.F., S. Sivapalaratnam and M.A.L reviewed the paper; W.H.O. conceived and managed the project, wrote the paper and chaired of the Genomics in Thrombosis and Haemostasis ISTH-SCC Committee; E.T. supervised the data analysis and wrote the paper.

Conflict-of-interest disclosure

The authors declare no competing financial interests.

REFERENCES

1. Shen T, Pajaro-Van de Stadt SH, Yeat NC, Lin JC. Clinical applications of next generation sequencing in cancer: from panels, to exomes, to genomes. *Front Genet*. 2015;6:215.
2. Ng PC, Kirkness EF. Whole genome sequencing. *Methods Mol Biol*. 2010;628:215-226.
3. Bodi K, Perera AG, Adams PS, et al. Comparison of commercially available target enrichment methods for next-generation sequencing. *J Biomol Tech*. 2013;24(2):73-86.
4. Westbury SK, Turro E, Greene D, et al. Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Med*. 2015;7(1):36.
5. Franchini M, Mannucci PM. Past, present and future of hemophilia: a narrative review. *Orphanet J Rare Dis*. 2012;7:24.
6. James PD, Goodeve AC. von Willebrand disease. *Genet Med*. 2011;13(5):365-376.
7. Laffan MA, Lester W, O'Donnell JS, et al. The diagnosis and management of von Willebrand disease: a United Kingdom Haemophilia Centre Doctors Organization guideline approved by the British Committee for Standards in Haematology. *Br J Haematol*. 2014;167(4):453-465.
8. Gresele P, Harrison P, Bury L, et al. Diagnosis of suspected inherited platelet function disorders: results of a worldwide survey. *J Thromb Haemost*. 2014;12(9):1562-1569.
9. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
10. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.
11. Plagnol V, Curtis J, Epstein M, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*. 2012;28(21):2747-2754.
12. International HapMap C, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010;467(7311):52-58.
13. Genomes Project C, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65.
14. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575.
15. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;6(2):80-92.
16. Stenson PD, Mort M, Ball EV, et al. The Human Gene Mutation Database: 2008 update. *Genome Med*. 2009;1(1):13.

17. Albers CA, Newbury-Ecob R, Ouwehand WH, Ghevaert C. New insights into the genetic basis of TAR (thrombocytopenia-absent radii) syndrome. *Curr Opin Genet Dev*. 2013;23(3):316-323.
18. Albers CA, Paul DS, Schulze H, et al. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nat Genet*. 2012;44(4):435-439, S431-432.
19. Johnsen JM, Auer PL, Morrison AC, et al. Common and rare von Willebrand factor (VWF) coding variants, VWF levels, and factor VIII levels in African Americans: the NHLBI Exome Sequencing Project. *Blood*. 2013;122(4):590-597.
20. Bellissimo DB, Christopherson PA, Flood VH, et al. VWF mutations and new sequence variations identified in healthy controls are more frequent in the African-American population. *Blood*. 2012;119(9):2135-2140.
21. Wang QY, Song J, Gibbs RA, Boerwinkle E, Dong JF, Yu FL. Characterizing polymorphisms and allelic diversity of von Willebrand factor gene in the 1000 Genomes. *J Thromb Haemost*. 2013;11(2):261-269.
22. Kujovich JL. Factor V Leiden thrombophilia. *Genet Med*. 2011;13(1):1-16.
23. Robinson PN, Kohler S, Oellrich A, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res*. 2014;24(2):340-348.
24. Lin D. An Information-Theoretic Definition of Similarity. *Proceedings of the Fifteenth International Conference on Machine Learning*. 1998;Morgan Kaufmann Publishers Inc., San Francisco, CA, USA:296-304.
25. Consortium UK, Walter K, Min JL, et al. The UK10K project identifies rare variants in health and disease. *Nature*. 2015;526(7571):82-90.
26. Matthijs G, Souche E, Alders M, et al. Guidelines for diagnostic next-generation sequencing. *Eur J Hum Genet*. 2015.
27. Gresele P, Subcommittee on Platelet P. Diagnosis of inherited platelet function disorders: guidance from the SSC of the ISTH. *J Thromb Haemost*. 2015;13(2):314-322.
28. Hayward CP, Rivard GE. Quebec platelet disorder. *Expert Rev Hematol*. 2011;4(2):137-141.
29. Eikenboom J, Van Marion V, Putter H, et al. Linkage analysis in families diagnosed with type 1 von Willebrand disease in the European study, molecular and clinical markers for the diagnosis and management of type 1 VWD. *J Thromb Haemost*. 2006;4(4):774-782.
30. Lutskiy MI, Rosen FS, Remold-O'Donnell E. Genotype-proteotype linkage in the Wiskott-Aldrich syndrome. *J Immunol*. 2005;175(2):1329-1336.
31. Jenny RJ, Pittman DD, Toole JJ, et al. Complete cDNA and derived amino acid sequence of human factor V. *Proc Natl Acad Sci U S A*. 1987;84(14):4846-4850.
32. Ajzner EE, Balogh I, Szabo T, Marosi A, Haramura G, Muszbek L. Severe coagulation factor V deficiency caused by 2 novel frameshift mutations: 2952delT in exon 13 and 5493insG in exon 16 of factor 5 gene. *Blood*. 2002;99(2):702-705.

33. Gitschier J, Wood WI, Goralka TM, et al. Characterization of the human factor VIII gene. *Nature*. 1984;312(5992):326-330.
34. Gitschier J, Wood WI, Tuddenham EG, et al. Detection and sequence of mutations in the factor VIII gene of haemophiliacs. *Nature*. 1985;315(6018):427-430.
35. Morange PE, Suchon P, Tregouet DA. Genetics of Venous Thrombosis: update in 2015. *Thromb Haemost*. 2015;114(5):910-919.
36. Nurden P, Debili N, Coupry I, et al. Thrombocytopenia resulting from mutations in filamin A can be expressed as an isolated syndrome. *Blood*. 2011;118(22):5928-5937.
37. Dargaud Y, Scoazec JY, Wielders SJ, et al. Characterization of an autosomal dominant bleeding disorder caused by a thrombomodulin mutation. *Blood*. 2015;125(9):1497-1501.
38. Langdown J, Luddington RJ, Huntington JA, Baglin TP. A hereditary bleeding disorder resulting from a premature stop codon in thrombomodulin (p.Cys537Stop). *Blood*. 2014;124(12):1951-1956.
39. Keeney S, Collins P, Cumming A, Goodeve A, Pasi J. Diagnosis and management of von Willebrand disease in the United Kingdom. *Semin Thromb Hemost*. 2011;37(5):488-494.
40. Derry JM, Ochs HD, Francke U. Isolation of a novel gene mutated in Wiskott-Aldrich syndrome. *Cell*. 1994;78(4):635-644.
41. Provaznikova D, Geierova V, Kumstyrova T, et al. Clinical manifestation and molecular genetic characterization of MYH9 disorders. *Platelets*. 2009;20(5):289-296.
42. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res*. 2002;30(12):e57.

Tables

Table 1. The 63 BPD genes present in the ThromboGenomics platform. BPDs targeted by the ThromboGenomics platform, grouped by disorder type and gene. For each gene and disease, the main mode of inheritance (MOI — AR: autosomal recessive; AD: autosomal dominant. XR: X-linked recessive) and the number of individuals in the *known*, *suspected* and *uncertain* categories found to carry a pathogenic variant by the ThromboGenomics platform are shown, with sub-totals for each set of disorders shown in brackets. One patient in the *uncertain* group is shown on two rows because she was given a digenic molecular diagnosis involving a likely pathogenic variant in *SERPINC1* and another in *PROC. GP1BA* appears twice because variants therein may be implicated in disorders listed on two separate rows (Bernard-Soulier syndrome and platelet-type von Willebrand disease). Note that gain-of-function variants in coagulation factor genes *F2*, *F5*, *FGA*, *FGB*, *FGG* may be involved in thrombotic disorders but these are not shown, with the exception of Factor V Leiden.

Coagulation Factor Disorders	**Genes**	**Main MOI**	***known* (41)**	***suspected* (26)**	***uncertain* (2)**
Combined V and VIII deficiency	*LMAN1*; *MCFD2*	AR; AR			
Factor V deficiency and Factor V Leiden	*F5*	AR	4	5	1
Factor VII deficiency	*F7*	AR	4	5	
Factor X deficiency	*F10*	AR	2		
Factor XI deficiency	*F11*	AR	2	1	1
Factor XIII deficiency	*F13A1*; *F13B*	AR; AR	2; 0	1; 0	
Fibrinogen deficiency	*FGA*; *FGB*; *FGG*	AD; AR; AR	7; 1; 2	3; 0; 0	
Fletcher factor deficiency	*KLKB1*	AR			
Haemophilia A	*F8*	XR	5	3	0

Haemophilia B	*F9*	XR	5		
Multiple coagulation factor deficiency type 1	*GGCX*	AR			
Multiple coagulation factor deficiency type 2	*VKORC1*	AR			
Plasminogen activator inhibitor 1 deficiency	*SERPINE1*	AR			
Plasminogen deficiency	*PLG*	AR		1	
Prothrombin deficiency	*F2*	AR	2	4	
Kininogen deficiency	*KNG1*	AR			
von Willebrand disease	*VWF*	AD	5	3	

Platelet Disorders	**Genes**	**Main MOI**	***known* (110)**	***suspected* (23)**	***uncertain* (4)**
ADP receptor defect	*P2RY12*	AR	1		
Amegakaryocytic thrombocytopenia with radioulnar synostosis	*HOXA11*	AD			
Autosomal dominant thrombocytopenia	*ANKRD26*; *CYCS*; *TUBB1*	AD; AD; AD	1; 4; 5	0; 0; 1	0; 0; 1
Bernard-Soulier syndrome	*GP1BA*; *GP1BB*; *GP9*	AR; AR; AR	3; 1; 0	0; 1; 0	
Bleeding diathesis due to glycoprotein VI deficiency	*GP6*	AR	1		
Chediak-Higashi syndrome	*LYST*	AR			
Congenital amegakaryocytic thrombocytopenia (CAMT)	*MPL*	AR	5		
Cyclic thrombocytopenia and thrombocythemia 1	*THPO*	AD			
Deficiency of phospholipase A2, group IVA	*PLA2G4A*	AR			
Dense granule abnormalities	*NBEA*	AD			
Familial platelet disorder with predisposition to AML	*RUNX1*	AD	2		2

Filamin A-related disorders	*FLNA*	XR	1		
Ghosal syndrome	*TBXAS1*	AR			
Glanzmann thrombasthenia	*ITGA2B*; *ITGB3*	AR; AR	5; 6	4; 0	
Gray platelet syndrome	*NBEAL2*	AR	8		
Hermansky-Pudlak syndrome	*HPS1*; *AP3B1*; *HPS3*; *HPS4*; *HPS5*; *HPS6*; *DTNBP1*; *BLOC1S3*	AR; AR; AR; AR; AR; AR; AR; AR	1; 0; 2; 0; 0; 2; 1; 0	4; 0; 0; 0; 2; 3; 0; 0	
May-Hegglin and other *MYH9*-related disorders	*MYH9*	AD	19	4	1
Paris-Trousseau thrombocytopenia and Jacobson syndrome	*FLI1*	AD			
Platelet-type von Willebrand disease	*GP1BA*	AD	10		
Québec platelet disorder	*PLAU*	AD		2	
Thrombocytopenia absent radius (TAR) syndrome	*RBM8A*	AR	10		
Thromboxane A2 receptor defect	*TBXA2R*	AR	2	1	
Wiskott-Aldrich syndrome	*WAS*	XR	16	1	
X-linked thrombocytopenia with dyserythropoiesis	*GATA1*	XR	4		

Thrombotic Disorders	**Genes**	**Main MOI**	***known* (8)**	***suspected* (7)**	***uncertain* (3)**
Antithrombin deficiency	*SERPINC1*	AR	4		1
Heparin co-factor 2 deficiency	*SERPIND1*	AD			
Histidine-rich glycoprotein deficiency	*HRG*	AD			
Protein C deficiency	*PROC*	AR	2	3	1
Protein S deficiency	*PROS1*	AR	2	4	1
Thrombomodulin deficiency	*THBD*	AD			
Tissue plasminogen activator deficiency	*PLAT*	AD			

Figure Legends

Figure 1. Breakdown of the 300 samples sequenced with the ThromboGenomics platform. The width of each box is proportional to the number of individuals it represents. The four main categories are shown as labels in italics. The shaded area in each box reflects the proportion of samples in which pathogenic or likely pathogenic variants were identified with the ThromboGenomics platform. Note that the mother of a haemophilia A patient from the *suspected* group appears in shading in the box representing the *unaffected* group.

Figure 2. Technical evaluation of the ThromboGenomics platform. A. Histogram of mean autosomal target coverage for 321 samples. **B.** Mean fraction of exonic (solid black) bases and HGMD variants (dashed red) covered at least at 0X, 1X, …, 50X. **C.** Coverage profile for the *ITGA2B* gene.

Figure 3. Sample identity assurance. A. The het/hom ratio versus the aut/X ratio is used to infer the gender of each individual. One sample from a male individual with an abnormally high aut/X ratio was substantially more degraded than all others. **B.** A scatterplot of the first two principal components derived from the 1000 Genomes genotypes, with individuals coloured by major population and projected ThromboGenomics individuals shown as black circles if they have fewer than seven candidate variants and triangles if they have at least seven candidate variants. For clarity, admixed American HapMap individuals are not shown. There is a lower density of ThromboGenomics individuals with African or East Asian ancestry but they all have at least seven variants, while approximately 80% of ThromboGenomics individuals with European ancestry have fewer than seven variants.

Figure 4. Candidate variants per sample. A-**C.** Bar plots of the number of candidate SNVs, indels and CNVs per individual. **D.** Scatterplot of the Bayes

Factor versus the observed over expected reads ratio for each CNV called by ExomeDepth and the thresholds distinguishing different levels of changes in zygosity. Note that the number of called CNVs is slightly biased upwards relative to the number of true CNVs because a single underlying CNV can sometimes be coded as multiple adjacent calls by the ExomeDepth algorithm. The fraction of CNVs surviving filtering is slightly elevated relative to the fraction of indels because we include calls with a Bayes factor down to 4.5 for maximum sensitivity and because they do not undergo any external cohort-based frequency filtering.

Figure 5. Case study. A. Human Phenotype Ontology (HPO) encoded phenotype of a case in the *suspected* category, visualised as a graph using the hpoPlot package. Note that “Abnormality of leukocytes” is also an “Abnormality of the immune system” (not shown). **B.** The ratio between observed and expected read depth over the *PLAU* gene for the case is shown in red and superimposed over the 95% confidence interval shown as a grey shaded area. In the lower panel the central position of each exon of the *PLAU* gene is shown as a vertical bar and the gene coordinates are provided on the horizontal axis. The data indicate that the case carries an additional copy of the *PLAU* gene (Bayes factor = 145), which is compatible with a diagnosis of suspected Québec platelet syndrome.

Figure 6. HPO-based prioritisation. A. HPO profile of a case with BSS encoded as a graph. Note atypical presence of hearing impairment, which is likely unrelated to the BSS. The plot beneath the graph shows the similarities between the patient profile and each gene in which the case has a candidate variant. The profile of *GP1BB* is the most similar out of the four genes with candidate variants. **B**. For each of the 109 HPO-coded cases for which a causative variant was assigned by the MDT, the similarity is shown between the case profile and the profiles of the genes in which the case has a candidate variant. The similarity to the gene containing the variant(s)

determined to be pathogenic or likely pathogenic in each case is shown as a red circle and the similarity to other genes containing variants of unknown significance are shown as gray dashes. Case index 1 corresponds to the BSS case shown in **A**.

Figure 7. HGMD variants and corresponding minor allele frequencies in ExAC. A. Truncated log-scale barplot showing the number of HGMD variants by HGMD phenotype. **B.** Log-scale barplot showing the number of HGMD variants binned by MAF in ExAC. **C.** Histogram of the 140 variants in BPD genes with a MAF in ExAC exceeding 1/1,000 (i.e. belonging to the blue bins in panel **B**) broken down by HGMD phenotype. The individual Phred-scaled MAFs of the variants (i.e. such that 30 corresponds to 1/1,000 and 20 to 1/100) are superimposed on the histogram and coloured by whether they are classified as disease causing (DM and DM? categories).

Figures

Figure 1

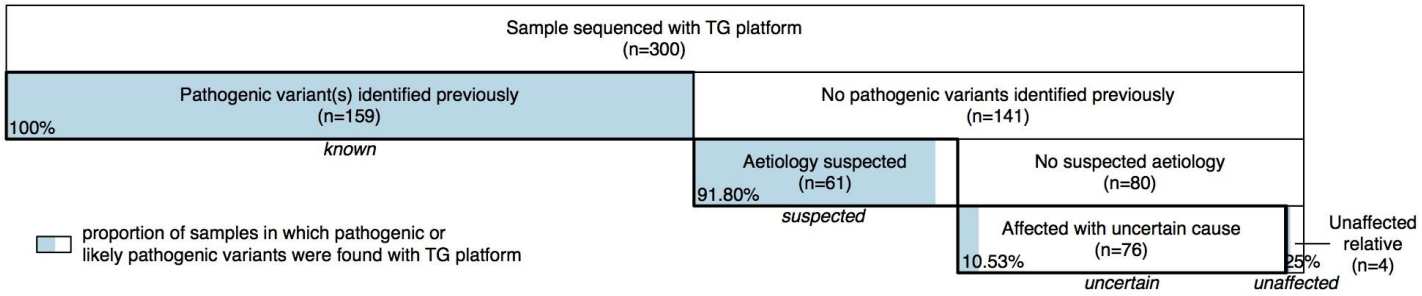


Figure 2

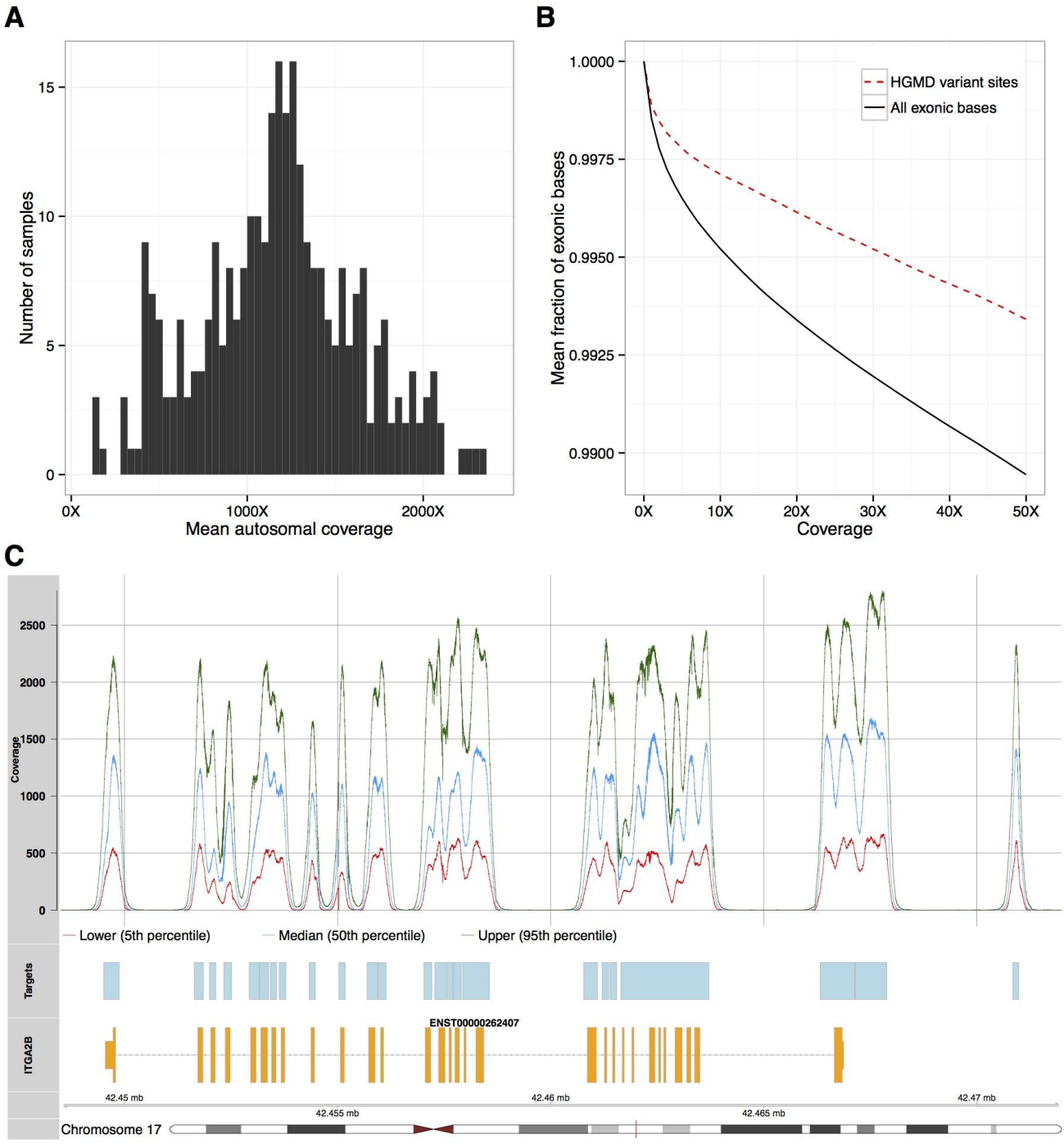

A
Number of samples
15
10
5
0
0X
1000X
2000X
Mean autosomal coverage
B
Mean fraction of exonic bases
1.0000
0.9975
0.9950
0.9925
0.9900
HGMD variant sites
All exonic bases
0X
10X
20X
30X
40X
50X
Coverage
C
Coverage
2500
2000
1500
1000
500
0
Lower (5th percentile)
Median (50th percentile)
Upper (95th percentile)
Targets
ITGA2B
ENST00000262407
42.45 mb
42.455 mb
42.46 mb
42.465 mb
42.47 mb
Chromosome 17

Figure 3

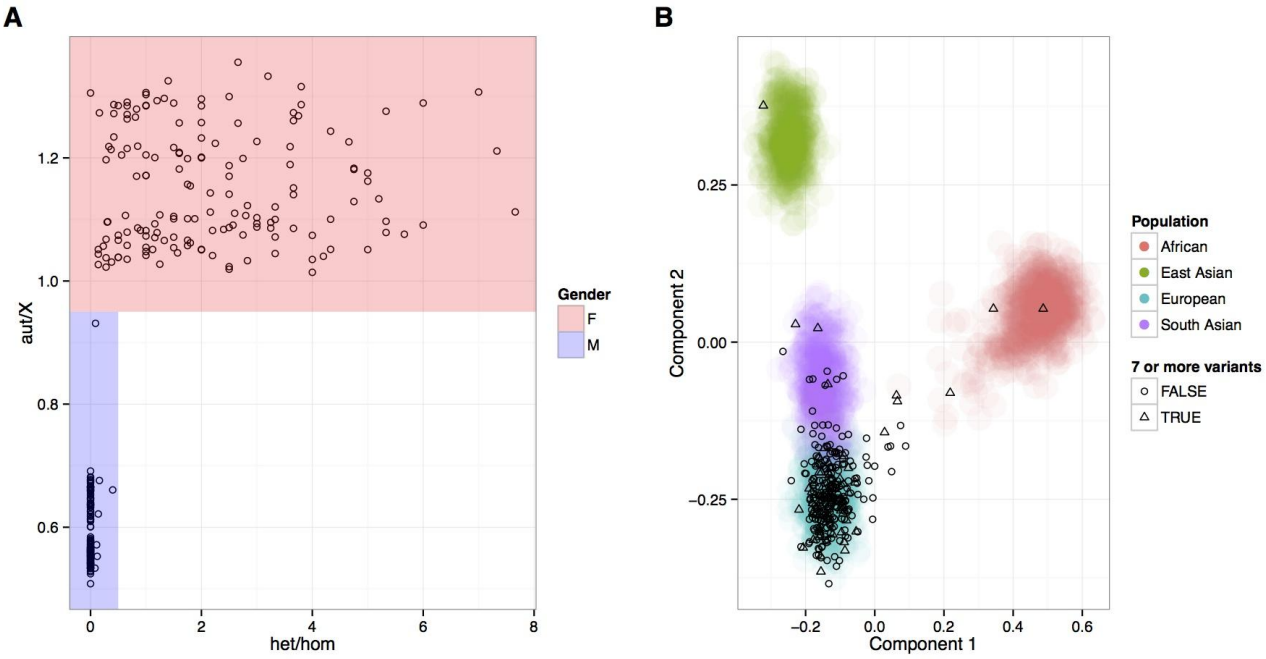

A
aut/X
het/hom
Gender
F
M
B
Component 2
Component 1
Population
African
East Asian
European
South Asian
7 or more variants
FALSE
TRUE

Figure 4

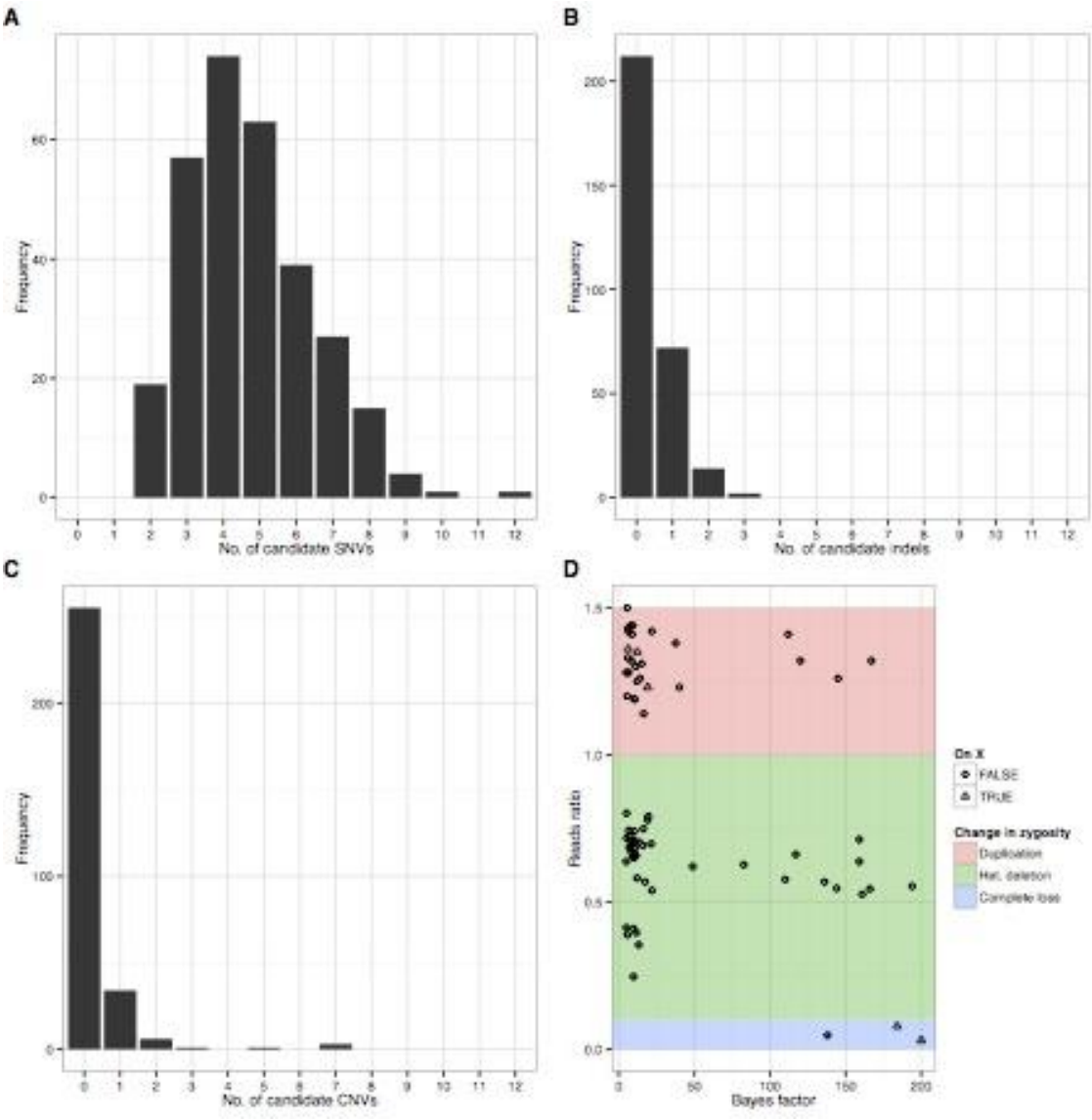

A
Frequency
No. of candidate SNVs
B
Frequency
No. of candidate indels
C
Frequency
No. of candidate CNVs
D
Reads ratio
Bayes factor
On X
FALSE
TRUE
Change in zygosity
Duplication
Het. deletion
Complete loss

Figure 5

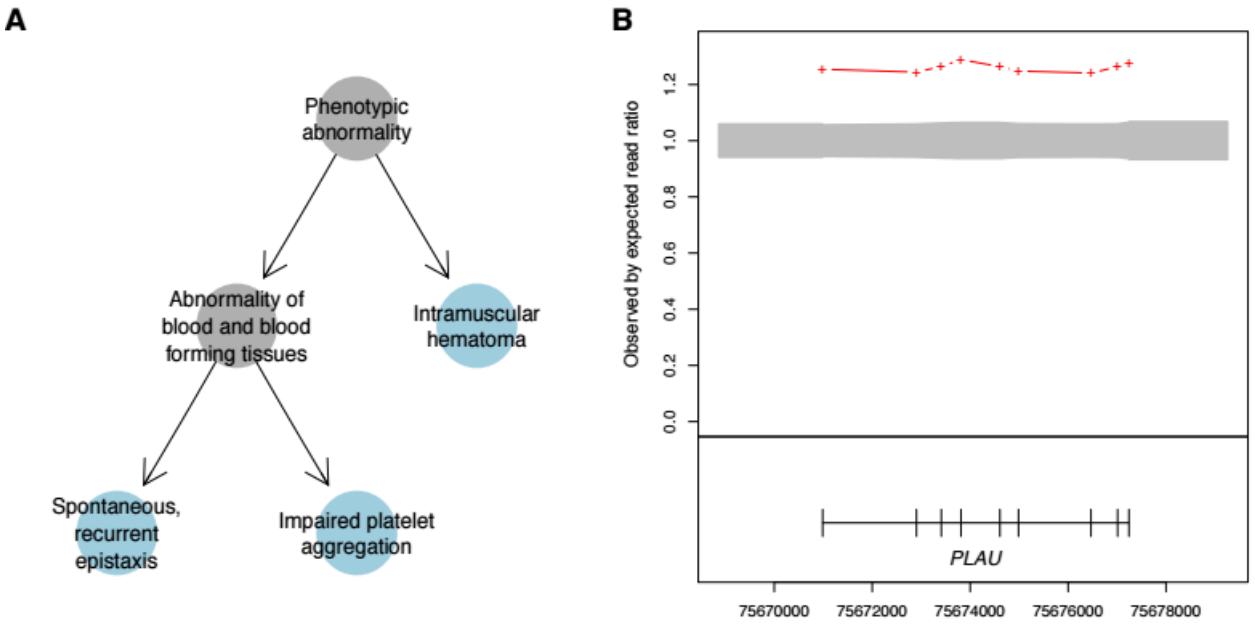

A
Phenotypic abnormality
Abnormality of blood and blood forming tissues
Intramuscular hematoma
Spontaneous, recurrent epistaxis
Impaired platelet aggregation
B
Observed by expected read ratio
0.0
0.2
0.4
0.6
0.8
1.0
1.2
PLAU
75670000
75672000
75674000
75676000
75678000

Figure 6

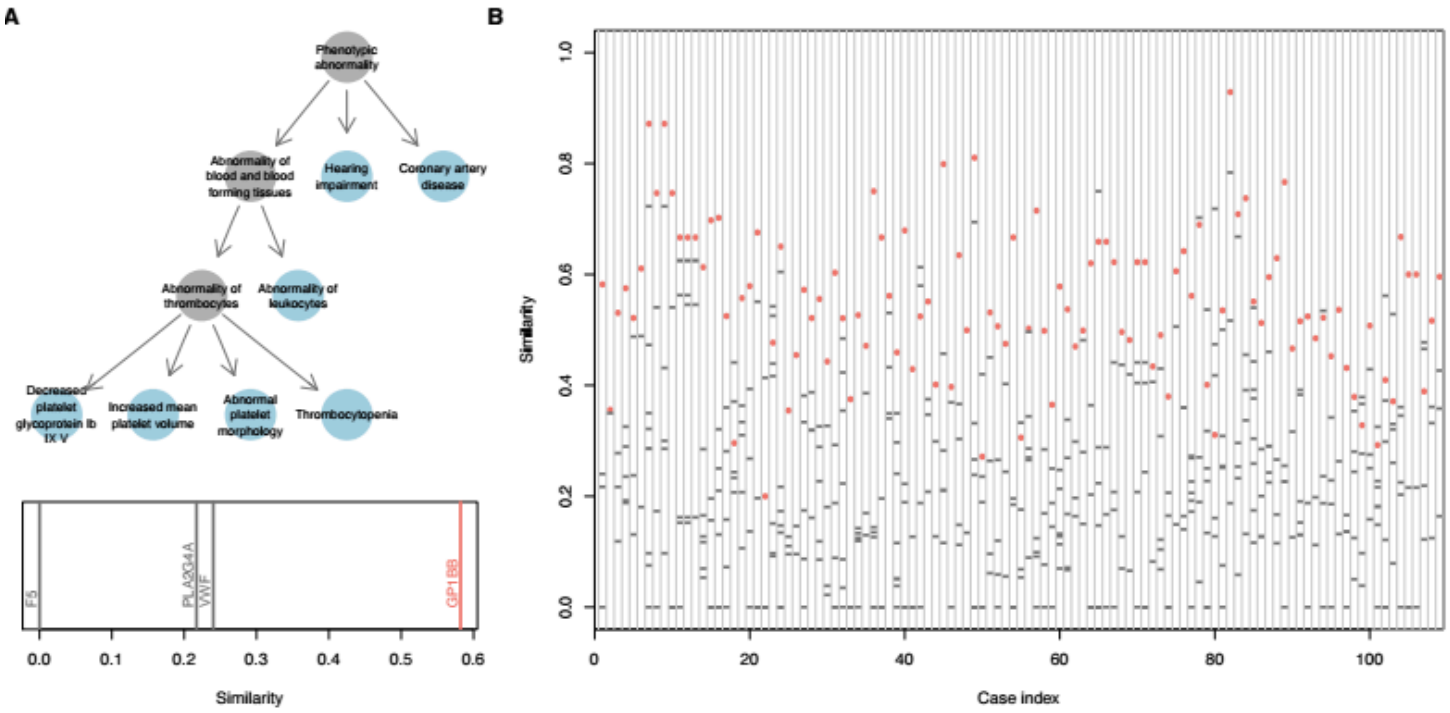
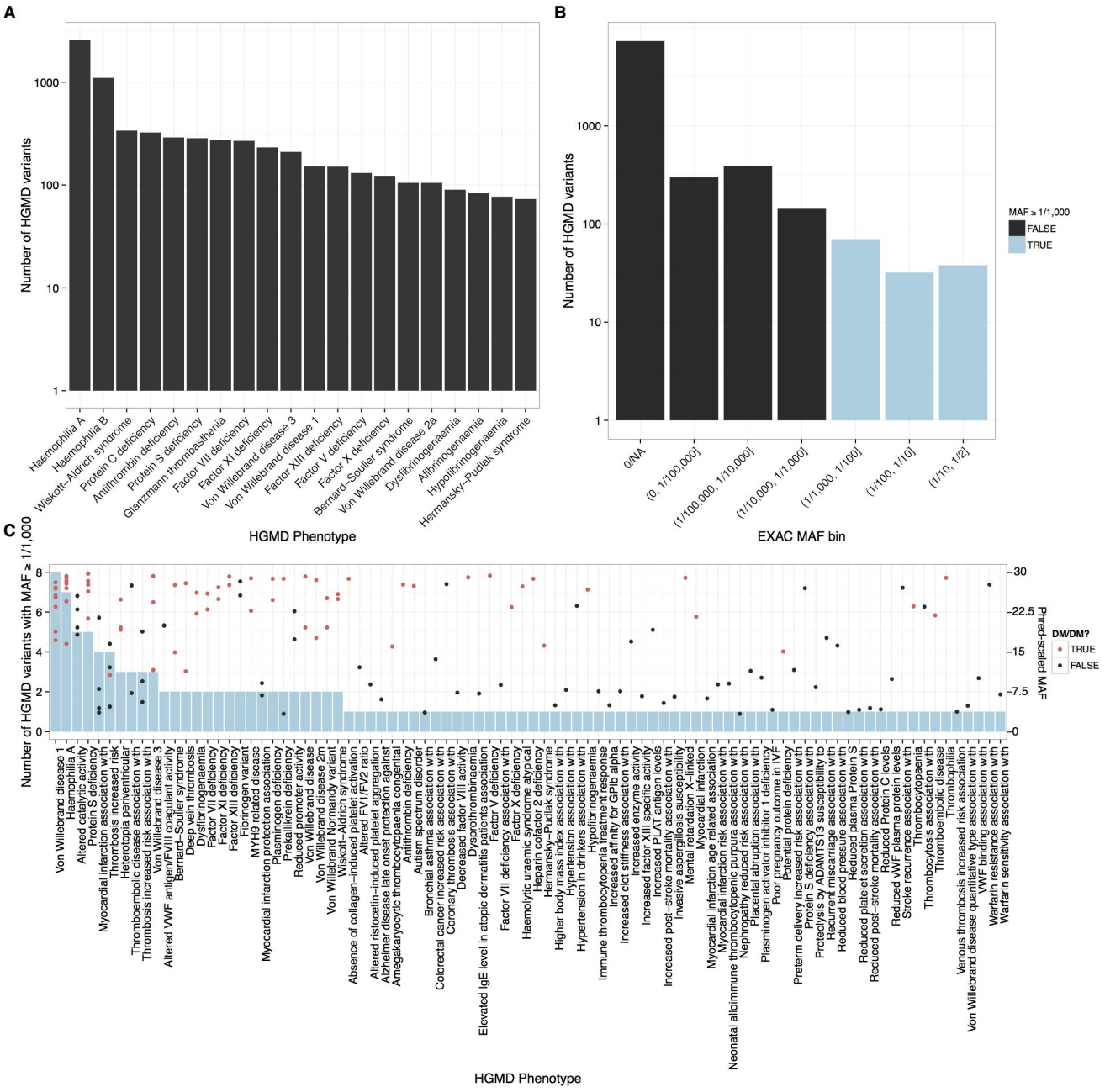

A
Phenotypic abnormality
Abnormality of blood and blood forming tissues
Hearing impairment
Coronary artery disease
Abnormality of thrombocytes
Abnormality of leukocytes
Decreased platelet glycoprotein Ib IX V
Increased mean platelet volume
Abnormal platelet morphology
Thrombocytopenia
F5
PLA2G4A
VWF
GP1BB
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Similarity
B
Similarity
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Case index

Figure 7


A
Number of HGMD variants
1000
100
10
1
Haemophilia A
Haemophilia B
Wiskott–Aldrich syndrome
Protein C deficiency
Antithrombin deficiency
Protein S deficiency
Glanzmann thrombasthenia
Factor VII deficiency
Factor XI deficiency
Von Willebrand disease 3
Von Willebrand disease 1
Factor XIII deficiency
Factor V deficiency
Factor X deficiency
Bernard–Soulier syndrome
Von Willebrand disease 2a
Dysfibrinogenaemia
Afibrinogenaemia
Hypofibrinogenaemia
Hermansky–Pudlak syndrome
HGMD Phenotype
B
Number of HGMD variants
1000
100
10
1
0/NA
(0, 1/100,000]
(1/100,000, 1/10,000]
(1/10,000, 1/1,000]
(1/1,000, 1/100]
(1/100, 1/10]
(1/10, 1/2]
EXAC MAF bin
MAF ≥ 1/1,000
FALSE
TRUE
C
Number of HGMD variants with MAF ≥ 1/1,000
8
6
4
2
0
30
22.5
15
7.5
0
Phred-scaled MAF
DM/DM?
TRUE
FALSE
Von Willebrand disease 1
Haemophilia A
Altered catalytic activity
Protein S deficiency
Myocardial infarction association with
Thrombosis increased risk
Heterotopia periventricular
Thromboembolic disease association with
Thrombosis increased risk association with
Von Willebrand disease 3
Altered VWF antigen/FVIII coagulant activity
Bernard–Soulier syndrome
Deep vein thrombosis
Dysfibrinogenaemia
Factor VII deficiency
Factor XI deficiency
Factor XIII deficiency
Fibrinogen variant
MYH9 related disease
Myocardial infarction protection association
Plasminogen deficiency
Prekallikrein deficiency
Reduced promoter activity
Von Willebrand disease
Von Willebrand disease 2m
Von Willebrand Normandy variant
Wiskott–Aldrich syndrome
Absence of collagen–induced platelet activation
Altered FVIII/FVIII:C ratio
Altered ristocetin–induced platelet agglutination
Alzheimer disease late onset protection against
Amegakaryocytic thrombocytopaenia congenital
Antithrombin deficiency
Autism spectrum disorder
Bronchial asthma association with
Colorectal cancer increased risk association with
Coronary thrombosis association with
Decreased factor VIII activity
Dysprothrombinaemia
Elevated IgE level in atopic dermatitis patients association
Factor V deficiency
Factor VII deficiency association with
Factor X deficiency
Haemolytic uraemic syndrome atypical
Heparin cofactor 2 deficiency
Hermansky–Pudlak syndrome
Higher body mass index association with
Hypertension association with
Hypertension in drinkers association with
Hypofibrinogenaemia
Immune thrombocytopenia treatment response
Increased affinity for GPIb alpha
Increased clot stiffness association with
Increased enzyme activity
Increased factor XI levels
Increased PLAT specific levels
Increased post–stroke mortality association with
Invasive aspergillosis susceptibility
Mental retardation X–linked
Myocardial infarction
Myocardial infarction age related association
Myocardial infarction risk association with
Neonatal alloimmune thrombocytopenic purpura association with
Nephropathy reduced risk association with
Placental abruption association with
Plasminogen activator inhibitor 1 deficiency
Poor pregnancy outcome in IVF
Preeclampsia risk association with
Preterm delivery increased risk association with
Protein S deficiency association with
Proteolysis by ADAMTS13 susceptibility to
Recurrent miscarriage association with
Reduced blood pressure association with
Reduced plasma Protein S
Reduced platelet secretion association with
Reduced post–stroke mortality association with
Reduced Protein C levels
Reduced vWF plasma protein levels
Stroke recurrence association with
Thrombocytopaenia
Thrombocytosis association with
Thromboembolic disease
Thrombophilia
Venous thrombosis increased risk association
Von Willebrand disease quantitative type association with
VWF binding association with
Warfarin resistance association with
Warfarin sensitivity association with
HGMD Phenotype